Weighted Hierarchical Mixed Topological Map: une méthode de classification hiérarchique à deux niveaux

Mory Ouattara, Ndèye Niang, Sylvie Thiria, Corinne Mandin, Fouad Badran

Weighted Hierarchical Mixed Topological Map: une méthode de classification hiérarchique à deux niveaux basée sur les cartes topologiques mixtes pondérées

Mory Ouattara*,** Ndèye Niang**
Sylvie Thiria *** Corinne Mandin*
Fouad Badran **

*Centre Scientifique et Techniques du Bâtiment 84 Avenue Jean Jaurès-Champs sur Marne
77445 Marne la vallée Cedex 2
http://www.cstb.fr
mory.ouattara@cstb.fr, corinne.mandin@cstb.fr
**CNAM CEDRIC 292 rue St Martin 75141 France-Paris
http://www.cedric.cnam.fr
n-deye.niang_keita@cnam.fr
fouad.badran@cnam.fr
***LOCEAN Université Paris VI 4 place Jussieu, Tour 45-55 75252 PARIS Cedex05, FRANCE
thiria@locean-ipsl.upmc.fr

Résumé. Nous proposons une méthode de classification automatique d'individus décrits par des variables mixtes structurées en blocs. C'est une méthode hiérarchique à deux niveaux, semblable à l'Analyse en Composantes Principales Hiérarchique de Wold, basée sur les cartes topologiques mixtes (MTM). La première étape consiste à établir pour chaque bloc de variables mixtes une classification dite locale représentant une synthèse des individus au niveau du bloc. La deuxième consiste à appliquer MTM sur les résultats du premier niveau pondérés par des indices de qualité des cartes tels que l'indice de Davie-Bouldin. On obtient alors une unique classification globale des individus, consensus entre les partitions issues du niveau 1. La méthode proposée est illustrée sur les données de la campagne nationale «logements» de l'Observatoire de la Qualité de l'Air Intérieur ($OQAI$).

1 Introduction

Dans le domaine environnemental, les phénomènes entraînant la dégradation de l'environnement sont liés à diverses sources de pollution. La qualité de l'air peut par exemple être liée à l'action directe de l'homme sur son environnement ou à des phénomènes naturels. La diversité des sources et causes de pollution de l'air intérieur nécessite des données de très grande dimension qui engendrent en cas de modélisations mathématique et statistique la création de plusieurs blocs de variables. Ces blocs décrivent des phénomènes différents qui peuvent être

des sources de pollutions de l'air intérieur parfois identiques. Ces blocs sont généralement constitués de variables qualitatives et quantitatives, dites mixtes.

L'objectif est alors de quantifier dans un premier temps l'information à l'intérieur de chaque bloc. Dans un deuxième temps, il s'agira d'établir, grâce aux quantifications vectorielles obtenues au niveau de chaque bloc, un résumé final de l'information contenue dans l'ensemble des blocs par consensus entre les blocs.
La problématique de gestion des données mixtes est assez classique et de nombreuses méthodes ont été développées par différents auteurs pour y répondre.

Dans le cadre de l'analyse factorielle, l'étude de variables mixtes repose généralement soit sur un codagel, souvent disjonctif, des variables qualitatives pour ensuite les traiter comme des variables numériques à travers une analyse en composantes principales (ACP), soit sur une transformation adéquate des variables quantitatives en variables qualitatives (Escofier (1979)) en préalable à une analyse des correspondances multiples (ACM). Saporta (1990) propose de réaliser une ACP, avec une métrique judicieusement choisie, sur le tableau résultant de la juxtaposition des variables quantitatives réduites et des variables qualitatives codées sous forme disjonctive complète. Pagès (2004) regroupe les points de vue d'Escofier et de Saporta dans l'analyse factorielle de données mixtes ($AFDM$).

Dans le cadre de la classification automatique d'individus, lorsque ces derniers sont décrits par des données mixtes, une extension naturelle serait d'appliquer la classification sur les coordonnées factorielles aux composantes issues de l'$AFDM$.

Dans le cadre des méthodes neuronales, la classification des données mixtes par la méthode des cartes auto-organisées de Kohonen (1995) repose sur une adaptation de la fonction de coût classique initialement définie pour les variables quantitatives.

Lebbah et al. (2005) proposent la méthode MTM (Mixed Topological Map) pour les données mixtes. C'est une combinaison des algorithmes SOM (Kohonen (1995)) et binbatch (Lebbah et al. (2005)) qui est une adaptation de l'algorithme de Kohonen aux variables qualitatives. On peut aussi noter les travaux de Cottrell et al. (2003) qui proposent la méthode $KACM$ (Kohonen Analyse des Correspondances Multiples) qui consiste à appliquer l'algorithme de Kohonen sur les coordonnées factorielles issues de l'application d'une ACM sur les variables qualitatives.

Pour la classification d'un ensemble d'individus décrits par des variables mixtes structurées en blocs, Niang et al. (2011) proposent la méthode $HMTM$ (Hierarchical Mixed Topological Map) qui consiste à appliquer MTM d'abord sur chaque bloc de variables mixtes et ensuite sur la table composée des résultats des blocs pris séparément. Cette méthode permet de résumer l'information à la fois dans chaque bloc et l'information générale issue de l'ensemble des blocs.

Dans cette communication, nous proposons $Weighted - HMTM$ qui est une extension de $HMTM$ dans laquelle une étape supplémentaire de pondération est ajoutée comme dans l'approche d'ACP hiérarchique de Wold et Kettanah (1996).

Dans la section 2, après avoir rappelé le principe de MTM et la méthode $HMTM$, nous décrivons la méthode $Weighted - HMTM$. En section 3 seront présentés les résultats de l'application à une base de données réelles issue de la campagne nationale « logements » de l'Observatoire de la Qualité de l'Air Intérieur organisée entre 2003 et 2005.

2 Méthodes de classification automatique d'individus décrits par des variables mixtes structurées en blocs

2.1 HMTM (Hierarchical Mixed Topological Map)

$HMTM$ consiste en une double application de l'algorithme MTM que nous décrivons ci-dessous. MTM est une extension de SOM aux données mixtes. L'algorithme MTM est une combinaison de l'algorithme de Kohonen (1995) pour les variables quantitatives et de l'algorithme Binbatch, pour les variables qualitatives codées sous la forme disjonctive (Lebbah et al., 2005). Ainsi, chaque individu z_i de la base de données E=$\{z_i, i = 1, ...N\}$ de dimension n comporte une partie quantitative représentée par z_i^r=$(z_{i1}...z_{ip})(z_i^r \in R^p)$ et une partie qualitative z_i^b=$(z_{i(p+1)}...z_{i(n)})$ $(z_i^b \in \{0,1\}^{n-p})$.
Comme dans SOM, on associe alors à chaque cellule c de la carte C un vecteur référent w_c qui se décompose en $w_c = (w_c^r, w_c^b)$ où $w_c^r \in R^p$ désigne la partie réelle du référent et $w_c^b \in R^{n-p}$ la partie binaire. Ainsi, le vecteur référent a la même structure que les données initiales.

On note par W l'ensemble des vecteurs référents, W^r sa partie réelle et W^b sa partie binaire. Nous décrivons ci-dessous les spécificités de la méthode MTM, en particulier la fonction de coût. Cette dernière repose sur une mesure D de dissimilarité entre un individu z_i et un référent w_c, elle est composée de la distance euclidienne pour la partie quantitative et de la distance de Hamming H pour la partie qualitative :

$$D(z_i, w_c) = ||z_i - w_c||^2 = ||z_i^r - w_c^r||^2 + \beta H(z_i^b, w_c^b) \tag{1}$$

où $\beta = \frac{p}{n-p}$ est le poids relatif des variables qualitatives par rapport aux variables quantitatives. La fonction de coût $J_{MTM}^T(\chi, w)$ à minimiser est alors :

$$J_{MTM}^T(\chi, w) = \sum_{z_i \in E} \sum_{c \in C} k^T(\sigma(\chi(z_i), c)) D(z_i, w_c) \tag{2}$$

où $\chi(z_i) = argmin_c(||z_i - w_c||^2), k^T(k > 0 \; et \; lim_{|x| \to 0}|k(x)| = 0)$, T et σ désignent respectivement la fonction d'affectation d'un individu au référent le plus proche, le système de voisinage associé à chaque référent, le paramètre de voisinage et le paramètre évaluant la distance en deux neurones. L'optimisation de la fonction de coût est réalisée itérativement de manière locale. Elle consiste à choisir le système de référents $w \in W$ qui minimise la fonction $J_{MTM}^T(\chi, w)$. Cette optimisation conduit aux expressions suivantes de calcul des référents :

$$w_c^r = \frac{\sum_{z_i \in E} k(\sigma(\chi(z_i), r)) z_i^r}{\sum_{z_i \in E} k(\sigma(\chi(z_i), r))} \tag{3}$$

pour la partie réelle et à

$$w_c^{bk} = \{_{1 \quad sinon}^{0 \; si \sum_{z_i \in E} k(\sigma(\chi(z_i),r))(1-z_i^{bk}) > \sum_{z_i \in E} k(\sigma(\chi(z_i),r))z_i^{bk}} \tag{4}$$

pour la partie binaire.

Le choix d'une carte topologique peut être effectué à l'aide de différents critères tels que l'erreur de quantification vectorielle, le taux d'erreur de classification topologique, la mesure de distorsion, les indices de Davies et Bouldin (1987) et la silhouette value de Rousseeuw (1979).
Dans le cas où la taille de la carte ou de manière équivalente le nombre de référents, est trop grand, il est possible d'utiliser une classification ascendante hiérarchique (CAH) pour regrouper les référents en un nombre restreint de classes (Yacoub et al., 2001).

Dans la méthode $HMTM$, Niang et al. (2011) ont appliqué dans une première étape MTM suivi d'une CAH sur les blocs de variables initiaux (Fig. 1) séparément. Cela fournit autant de partitions que de blocs. Ces partitions assimilables à des variables qualitatives sont regroupées dans une table (Fig. 2) sur laquelle on applique à nouveau MTM suivi d'une CAH. C'est donc une méthode à deux niveaux dans laquelle :

– Le niveau inférieur est constitué des blocs de variables initiaux (Fig. 1).

– Le niveau supérieur est constitué d'un bloc de variables catégorielles représentant la classe d'appartenance de chaque individu aux partitions correspondants aux blocs (table en bas à gauche de la Fig. 2).

La méthode $HMTM$ permet d'obtenir à la fois des typologies des individus propres aux variables de chaque bloc et une typologie globale des observations qui tient compte de l'information au niveau des blocs.

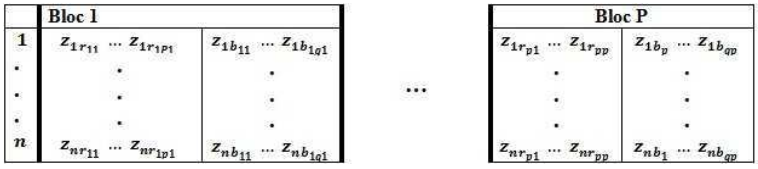

FIG. 1 – *Structure des tables au niveau inférieur*

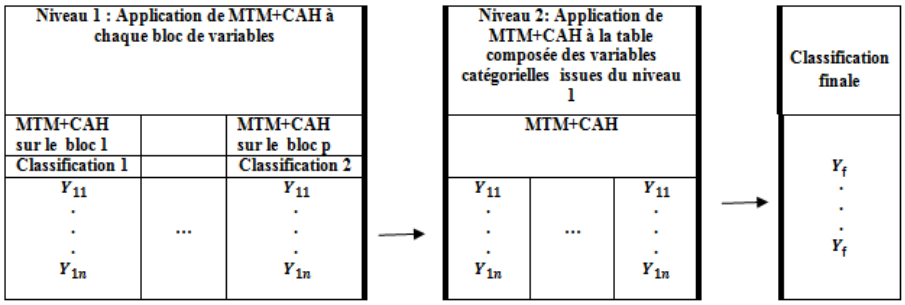

FIG. 2 – *Déroulement de l'algorithme HMTM.* Y_{ji} *est la modalité de la variable qualitative* Y_j *prise par l'individus i dans la classification obtenue au niveau du bloc j et* Y_f *la classification finale obtenue*

2.2 WEIGHTED-HMTM

Les différents indices d'appréciation de la qualité d'une classification montrent que les classifications obtenues au niveau de chaque bloc de variables mixtes sont de qualités différentes. Ils permettent par conséquent d'établir une hiérarchie de qualité entre les classifications. Nous proposons donc de tenir compte de cette hiérarchie en introduisant des coefficients de pondération dans la fonction de coût de $HMTM$. Cela consiste à remplaçer la distance euclidienne classique par une distance euclidienne pondérée sous la forme :

$$D^2(z_i, w_j) = \sum_{k=1}^{n} p_{ik}(z_{ik} - w_{jk})^2 \tag{5}$$

où w_j, z_i, p_{ik} et n correspondent respectivement au centre de classe ou référent, au i ième individu, au poids relatif des individus par rapport au référent w_j et à la dimension de l'espace. De plus, dans $weighted - HMTM$ ($WHMTM$) nous proposons de conserver l'information au niveau des référents dans la première étape.
Donc au niveau supérieur dans $WHMTM$ chaque individu i est représenté par un vecteur $V_i = (V_{i_{K_1}}, ..., V_{i_{K_b}})$ où $V_{i_{K_j}}$ est le vecteur référent ayant capté l'individu i dans le bloc j (cf. Fig 3). Les observations sont donc contenues dans un espace de dimension $(K_1 + ... + K_b)$ où b est le nombre de blocs et K_j la dimension du bloc j.

Ce choix permet de conserver plus d'information dans chaque bloc de variables.

Finalement, dans $WHMTM$, la fonction de coût associée aux données est :

$$J^T_{WHMTM}(\chi, \Omega) = \sum_{V_i \in E} \sum_{c \in C} k^T(\sigma(\chi(V_i), \Omega_c)) D^2(V_i, \Omega_c) \tag{6}$$

Où Ω_c est le référent au niveau supérieur. Formellement, p_{ik} est une fonction de la distance entre l'individu V_i et le référent de la carte C_l du bloc l auquel il appartient au niveau inférieur et de l'indice de Davies-Bouldin associé à cette carte.

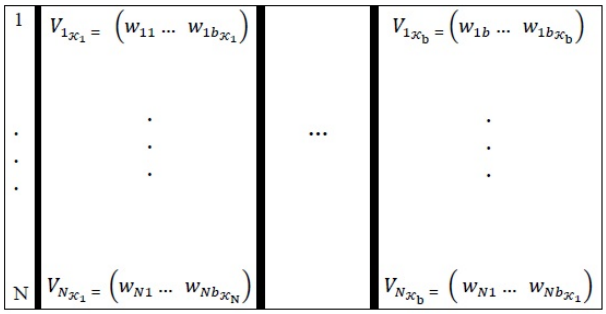

FIG. 3 – *Structure de la table au niveau supérieur dans WHMTM*

Les p_{ik} ne varient pas au cours de l'apprentissage au niveau supérieur et l'ensemble $\{p_{ik}, i \in 1, ..., N \; et \; k = 1, ..., b\}$ vérifie $\sum_i^n p_{ik} = 1 \; et \; 0 \leq p_{ik} \leq 1$. Sous ces hypothèses Huang et al. (2005) montrent que la fonction J^T_{WHMTM} converge. L'optimisation de cette fonction de coût conduit aux centres suivants pour un référent c :

la partie réelle est :

$$W_c^r = \frac{\sum_{V_i \in E} \sum_k p_{ik} k(\sigma(\chi(V_i), c)) V_i^r}{\sum_{V_i \in E} \sum_k p_{ik} k(\sigma(\chi(V_i), c))} \tag{7}$$

La composante k de la partie binaire est :

$$W_c^{bk} = \{^{0 \; si \sum_{V_i \in E} k(\sigma(\chi(V_i),c)) \sum_k p_{ik}(1-V_i^{bk}) > \sum_{V_i \in E} k(\sigma(\chi(V_i),c)) \sum_k p_{ik}(V_i^{bk})}_{1 sinon} \tag{8}$$

L'application de la CAH sur la carte finale fournit la partition d'individus consensus entre les blocs initiaux.

L'interprétation de cette partition consensus fait apparaître un ensemble de variables caractéristiques des classes. On aurait aussi pu considérer comme partition finale celle ayant le meilleur indice de Davies-Bouldin au niveau des blocs. Mais l'interprétation de cette partition en utilisant les variables du bloc associé et celles des autres blocs considérées comme variables illustratives engendrent une perte d'information quant aux variables effectives (variables de chaque bloc) qui caractérisent réellement les classifications dans ces blocs. Cela sera plus clairement illustrée dans l'application.

3 Application sur les données de la Campagne Nationale « Logements»

L'Observatoire de la Qualité de l'Air Intérieur (OQAI) a réalisé entre 2003 et 2005 une campagne nationale dans les logements sur un échantillon de 567 logements représentatif du parc des 24 millions de résidences principales de la France continentale métropolitaine. Cette

campagne vise à dresser un état de la pollution de l'air dans l'habitat afin de donner les éléments utiles pour l'estimation de l'exposition des populations, la quantification et la hiérarchisation des risques sanitaires associés, ainsi que l'identification des facteurs prédictifs de la qualité de l'air intérieur.
Des questions ont été posées sur : la structure du bâtiment (les matériaux utilisés pour la construction, les produits de décoration, le mobilier, ...), la structure des ménages et les habitudes des occupants.
Finalement une base de données composée de plus de 628 variables a été constituée. Cette base a été ensuite divisée en trois bases par regroupement et combinaison de variables suivant trois critères. Le premier critère décrit les logements, le deuxième critère décrit la structure des ménages et le troisième critère décrit les habitudes des occupants. Ces trois bases sont dans la suite appelées blocs de variables.

- Le bloc logements est composé de 71 variables dont 32 quantitatives.
- Le bloc ménages est composé de 12 variables dont 5 quantitatives.
- Le bloc habitudes est composé de 45 variables dont 21 quantitatives.

3.1 Apprentissage des cartes

Plusieurs apprentissages sont effectués au niveau des blocs de variables en faisant varier les paramètres de la fonction de coût, le nombre d'itérations n_i, les dimensions de la carte. La meilleure carte est alors choisie parmi l'ensemble des apprentissages, elle équivaut à la carte qui donne le meilleur indice de Davie Bouldin (Fig.4). Rappelons qu'il est possible d'utiliser d'autres indices.

3.2 Application de l'algorithme WHMTM

L'application de la méthode proposée sur chacun des blocs de variables de la campagne nationale « logements » montre à travers l'indice de Davie-Bouldin que la carte obtenue sur le bloc ménages est la meilleure au niveau inférieur. Dans la suite de cette section nous donnons d'abord une description des classes basée sur les variables les plus caratéristiques du bloc mais aussi à travers les variables des autres blocs considérées comme illustratives. Ensuite nous décrivons les classes obtenues après application de $WHMTM$ par rapport à l'ensemble des variables.

Au niveau inférieur :

Dans le bloc logement, on obtient une classification en 5 classes. Les variables les plus caractéristiques de ces classes sont : la surface du logement, l'année de construction du bâtiment, le type de logement (individuel ou collectif), l'ameublement intérieur des logements. Ainsi, les classes 1 et 2 contiennent les petits logements collectifs récents. La classe 3 regroupe les logements ou maisons récents de taille moyenne possédant beaucoup d'appareils à combustion raccordés à un conduit de fumée.
Les classes 4 et 5 regroupent les grandes maisons individuelles généralement tout en un raccordées à des conduits de fumée et possédant un fort taux de meubles en bois massifs (classe

4) et les petite maisons individuelles (classe 5).

En utilisant les variables des autres blocs pour enrichir l'interprétation on constate que seule la variable «revenu» du bloc ménage et certaines variables d'entretient du logement et de soins corporels du bloc habitudes permettent de caractériser les classes de logements. Par exemple les classes de logements 1 et 2 sont associées à des familles ayant de faibles revenus alors que les logements 3 et 4 sont associés à des familles ayant des revenus élevés. Notons que n'interviennent pas dans la description les variables décrivant le nombre d'enfants de moins ou de plus de 10 ans par foyer.

Dans le bloc ménage, on obtient une classification en 6 classes. Les variables les plus caractéristiques des classes sont : les revenus du ménage, le nombre de personnes, l'âge des enfants et le statut des ménages (personnes seules ou en couples).

En utilisant les variables des autres blocs on retrouve certaines variables significatives dans l'interprétation précédente comme la variable taux de meubles en bois du bloc logement et les variables d'entretient et de produits de soins corporels du bloc habitude qui permettent de caractériser les classes du bloc ménage. Par contre la variable âge du logement ne caractérise plus de manière significative les classes. De même l'âge des enfants significative ici ne l'était pas dans la partition du bloc logement. On constate alors une perte d'information dans les interprétations. Il en va de même pour l'interprétation de la partition associée au bloc habitude.

Au niveau supérieur, on obtient une classification en 8 classes après application d'une classification ascendante hiérarchique sur les neurones de ce niveau.

Les familles aisées, âgées, souvent nombreuses avec beaucoup d'enfants de plus ou de moins de 10 ans habitent généralement dans des maisons individuelles tout en un avec garage et cave attenants, ils entretiennent beaucoup leur maison. A contrario, dans les logements collectifs on retrouve les jeunes ou les retraités vivant seuls et ayant des revenus moyens, ils entretiennent peu leur logement. Le nombre d'enfants de plus ou de moins de 10 apparait important pour les classes de cette partition

En plus de cette confrontation au niveau de l'interprétation des classes, nous avons comparé les partitions en utilisant l'indice Rand qui met en évidence la similarité entre les partitions (TAB 1.), plus la valeur de cet indice est proche de 1 mieux est la similarité entre les deux classifications.

Les valeurs relativement élevées des indices de rand entre partitions du niveau inférieur montrent certes une similitude mais elle montre aussi le fait que chaque bloc de variables porte une information qui lui est spécifique. La méthode directe qui consiste à appliquer MTM sur la table juxtaposant les blocs initiaux (MTM_{gle}) fournit une partition globale moins semblable aux partitions locales à la différence des partitions obtenues avec $HMTM$ et $WHMTM$ plus consensuelles.

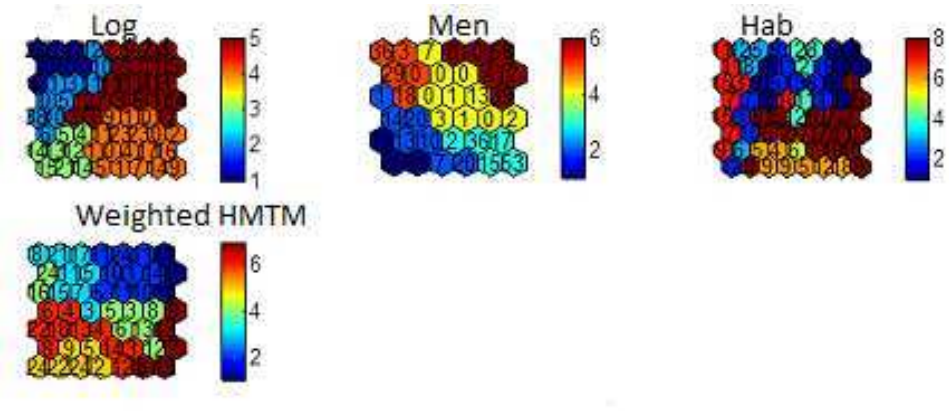

FIG. 4 – *Structure des cartes aux niveaux inférieur et supérieur : en haut les cartes du niveau inférieur et en bas la carte du niveau supérieur*

	Log	Men	Hab	weighted-HMTM$_{\Pi_k}$	MTM$_{gle}$
Log	1	0.69	0.70	0.82	0.73
Men		1	0.68	0.68	0.67
Hab			1	0.72	0.66
Weighted-HMTM$_{\Pi_k}$				1	70
MTM$_{gle}$					1

TAB. 1 – *Valeurs de l'indice de Rand de comparaison des classifications : Log, Men, Hab, weighted-HMTM$_{\Pi_k}$, MTM$_{gle}$ correspondent respectivement aux classifications obtenues sur les blocs logement, ménage, habitude, les individus sont affectés des poids P_{ik}, enfin la classification obtenue en concaténant les tables composées des données initiales dans chaque bloc*

4 Conclusion

Face à la problématique de classification d'individus décrits par des variables mixtes structurées en blocs, la méthode WHMTM crée un consensus entre l'information dans chaque bloc de variable mixtes et l'information globale issue de l'ensemble des variables. Les pondérations introduites au niveau supérieur permettent d'améliorer la structure de la carte finale. Cependant, le consensus peut être amélioré entre les blocs et nécessite des recherches supplémentaires. Notamment, en procédant à la sélection des variables les plus informatives dans le passage du niveau 1 au niveau 2, cela permettrait une réduction de la dimension des individus au niveau 2 et en rendant les poids adaptives.

Références

Cottrell, M., I. Smail, et P. Letrémy (2003). Cartes auto-organisés pour l'analyse exploratoire de données et la visualisation. *Journal de la Société Française de Statistique 144*, 67–106.

Davies, D. L. et D. W. Bouldin (1987). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence (2) 4*, 224–227.

Escofier, B. (1979). Traitement simultané de variables quantitatives et qualitatives en analyse factorielle. *RSA 4*, 137–146.

Huang, J. Z., M. K. N. H. Rong, et Z. LI (2005). Automated variable weighting in k-means type clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27*, 657–668.

Kohonen, T. (1995). the self organizing map.

Lebbah, M., A. Chazottes, F. Badran, et S. Thiria (2005). Cartes topologiques mixtes. *EGC 5*, 43–48.

Niang, N., M. Ouattara, J. Brajard, et C. Mandin (2011). classification d'individus decrits par des variables mixtes structurées en blocs. *SFDS 43*, 146–152.

Pagès (2004). *Analyse Factorielle de Données Mixtes*. ITALIA : RSA.

Rousseeuw, P. J. (1979). Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics, Volume 20, Page 65 20*, 53–65.

Saporta, G. (1990). Simultaneous analysis of qualitative and quantitative data atti della xxxv riunione scientifica. *Società itialana di statistica 35*, 57–64.

Wold, S. et N. Kettanah (1996). Hierarchical multiblock pls and pc models interpretation and as an aternative to variable selection. *J. Chemon 10*, 463–482.

Yacoub, M., N. Niang, F. Badran, et S. Thiria (2001). A new hierarchical clustering method using topological map. *ASMDA2001*.

Summary

We propose a method to solve the problem of clustering individuals described by mixed variables structured in blocks. It is a hierarchical method with two levels similar to Hierarchical Principal Component Analysis presented by Wold, based on the mixed topology maps (MTM). The first step consists in establishing, for each block of mixed variables, a clustering representing a synthesis of local individuals at the block level. The second step is to re-apply MTM on the weighted results of the first level. We obtain a global clustering of individuals, consensus among the partitions from level 1. The proposed method is illustrated on data from the national survey in the French housing stock carried out in 2003-2005 by the Observatory for Indoor Air Quality (OQAI).